Context-Aware Augmentation for Semantic Segmentation in Low-Data Regimes

Ariel Wang
Stanford University
ariwang@stanford.edu

Abstract

This paper presents a novel data augmentation technique for semantic segmentation in low-data regimes. The method generates out-of-distribution training samples by mixing semantic classes with controlled co-occurrence frequencies, thereby increasing contextual diversity in the dataset. A co-occurrence likelihood score enables precise control over how common or rare the class pairings are in the synthetic data, to adjust the similarity between the augmented dataset compared the the original data distribution. Two tunable hyper-parameters govern the frequency of combinations and the ratio of synthetic to real data. This approach improves pixel-wise segmentation accuracy in low-data regimes, increasing validation accuracy from 49.88% to 52.63% by augmenting the training set three-fold with synthetic images. These results show the importance of carefully controlling the distribution of augmented data in a low-data regime, as small datasets are prone to overfitting if overwhelmed with large amounts of synthetic data. This work highlights the potential of controlled, context-aware augmentation strategies to enhance semantic segmentation performance while mitigating the costs of manual labeling.

1. Introduction

Semantic segmentation datasets are significantly more expensive to produce than image classification datasets, as they require pixel-level annotations rather than simple image-level labels. For high-resolution datasets like ADE20K, which contain images commonly resized to 512×512 pixels, this labeling process demands intensive human labor to accurately delineate each object. This paper explores a data augmentation strategy that generates out-of-distribution training examples by overlaying objects from different semantic classes that rarely co-occur, as measured by a co-occurrence likelihood score. This approach serves two purposes: (1) it reduces the dependence on large, diverse labeled datasets by synthetically introducing novel contexts, and (2) it improves robustness to partial occlusions, encouraging the model to focus on class-specific visual cues rather than contextual priors. This paper uses a copy-paste technique to merge together features from different images as new images to the dataset. The paper outlines experimentation with two hyper-parameters that control the amount of augmented data compared to real data, and the uniqueness of the image pairs that generated the augmented data. By generating an augmented dataset 300% the size of the original dataset and adding it to the original dataset, the validation accuracy increased from 49.88% to 52.63% when evaluated on the same PSPNet model. The augmented data pairings were sourced from the rarest pairings of the original dataset. The higher the ratio of augmented data to real data, the more overfit the model became. When varying the rarity by selecting different quantiles of the co-occurrence list to use as the source of augmented pairs, any augmentation improves the model, but the more common pairings actually improved the model more compared to the rare pairings.

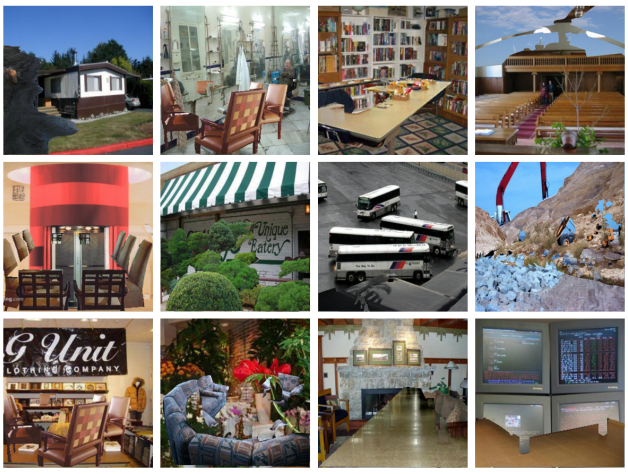

Figure 1. Examples of Synthetic Images using Augmentation Method

2. Related Work

Prior techniques such as ClassMix [2] and Copy-Paste [1] have demonstrated the benefits of mixing classes and

objects in segmentation tasks, primarily to enhance semi-supervised learning or increase data diversity. ClassMix, for instance, generates augmented blending labelled images together to enable more effective use of unlabeled data. This method relies on the assumption that combining similar regions (e.g., sky with road, or car with street) can help reinforce class boundaries and improve label propagation. Similarly, Copy-Paste focuses on object-level augmentation by extracting objects from source images and pasting them into new target scenes. This technique is typically used for semi-supervised models to improve classification when dealing with intra-class diversity and occlusion patterns. In contrast, our method introduces class combinations with controllable contextual rarity using a co-occurrence matrix. Rather than preserving only realistic compositions, we selectively introduce uncommon pairings to challenge the model's contextual assumptions. This is especially effective in low-data regimes—on the order of 1,000 images—where models tend to overfit to limited scene priors. By modulating contextual plausibility, our approach encourages better generalization to novel or out-of-distribution scenarios. In contrast, prior methods are typically applied in large-scale, semi-supervised settings where the inherent data diversity reduces the need for controlled mixing. In smaller-scale regimes, however—where training data may not reflect the test distribution—our ability to selectively inject augmentations based on co-occurrence likelihood allows for targeted diversification without overwhelming the dataset with synthetic examples.

3. Data

This project utilises the ADE20K dataset from MIT CSAIL, which includes over 25,000 semantic-segmentation labelled images and masks. The data set includes a diverse range of objects pictured both indoors and outdoors. The images are resized to be 3x512x512 for the training to simplify computation. Each image is accompanied by a segmentation mask that categorizes each pixel as 1 of 151 unique classes such as "person", "grass", "chair", etc. Below is an example of the raw images and segmentation masks.

This project tests a hypothesis that smaller datasets can be artificially augmented to improve learning objectives. After the data was separated into train, validation, and test sets, the training dataset was decreased to one sixteenth of its original size. The final training dataset consisted of 1010 images and the validation dataset consisted of 2000 images. To understand the distribution of the dataset, a histogram of pixel class labels were plotted. The most frequent class labels are background elements such as "wall" and "building," which tend to occupy large portions of the images. These classes dominate in terms of pixel coverage but exhibit relatively low variation in visual appearance

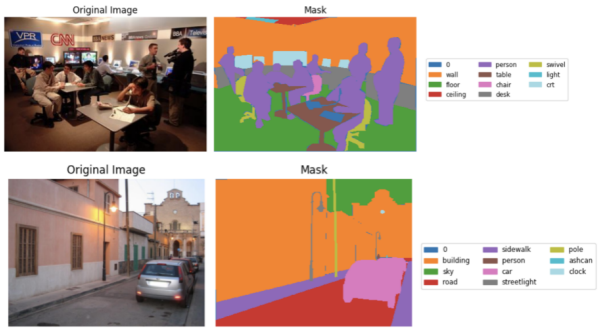

Figure 2. Example images from dataset; indoor and outdoor

across different scenes, and are likely to be easier to identify compared to intricate and highly variable features like "person". Furthermore, the 4th most ubiquitous label is "Class 0", which corresponds to unknown features in the dataset. Since there are no similarities between unknown pixels, "Class 0" was removed when calculating training loss to prevent overfitting and destabilizing the learning.

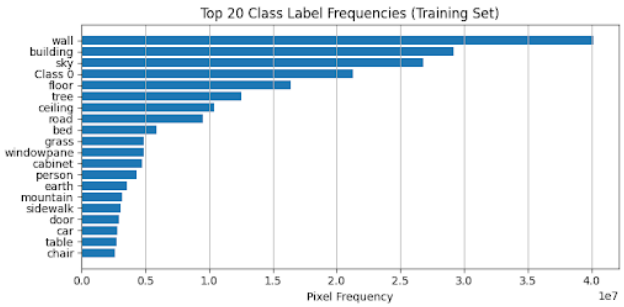

Figure 3. Pixel-wise frequency of each class label of training set

4. Methods

4.1. Architecture Choice

To test the hypothesis that context-clash augmentation improves segmentation models, a baseline model was first established. The architecture selected is a PSPNet (Pyramid Scene Parsing Network) with ResNet-18 as the backbone. This model is very standard for segmentation and has won the ImageNet Scene Parsing Challenge. It uses spatial pooling at multiple scales to capture features of different sizes. It then up-samples and concatenates these features with the original feature map, to enable the model to have global scene understanding with fine-grained details. Typically a PSPNet uses a ResNet-50 or ResNet-101 backbone. For ablation studies, both these models require too much memory and compute to run all the trials required to compare parameters. Thus a ResNet-18 backbone was subbed in, which has 18 layers instead of 50, resulting in 11.7 million parameters instead of 25.6 million. This prevented training

jobs from failing due to memory overflow or taking hours to train over thousands of high resolution images. The final architecture used ResNet-18 as the backbone, training dataset of 1010 images, batch size of 8 and 5 epochs for evaluation. A single training run took 8 minutes, which allowed for 3 repeated trials for each set of parameters to account for any weight initialization differences. All training and validation accuracies were averaged over 3 trials.

4.2. Pre-processing

The RGB images were all resized to be 512x512 pixels, then normalized using pre-computed mean and standard deviation from ImageNet. Instead of calculating the distribution over the training dataset, it saves compute to use values that are computed from 1.2 million images from ImageNet. By making all images the same size and normalizing, the training is more efficient because it does not need to account for different image sizes and models learn faster with zero mean and unit variance. When using a ResNet backbone it is also recommended to use ImageNet normalizing since that is how they are usually trained.

4.3. Training Setup

Semantic segmentation is a classification task at the pixel level, where each pixel is assigned to one of 151 classes. Therefore, this paper uses cross-entropy loss, which compares the predicted class probability distribution at each pixel with the ground truth label. This loss encourages the model to assign high confidence to the correct class for every pixel. The optimizer chosen is Adam, which is better suited to a more complex model like PSPnet with many different pooling layers. Thus, using Adam can make the model converge faster and be less sensitive to hyperparameters. The learning rate was tuned by overfitting a very small training to ensure the rate is not too low or high. Using a small dataset of 10, the learning rate of 0.001 was able to overfit to above 90%. The batch size of 8 was the highest my computer could handle without memory overflow, especially since the augmented datasets reached up to 5000+ high resolution images. Batch size, learning rate, optimizer were all kept constant for the ablation study.

4.4. Data Augmentation Technique

In order to augment the dataset by artificially creating context-clashing images, the dataset must first be analyzed to quantify the rarity of label combinations. Here, a list is used to compute the co-occurrence likelihood of each pair of labels. A helper function loops over each image in the dataset and if two labels both appear in an image the frequency counter for the pair iterates. At the end, all the frequencies are normalized to give a likelihood of a pair of labels occurring simultaneously. Context-associated pairs like "floor" with "ceiling" occur much more frequently than

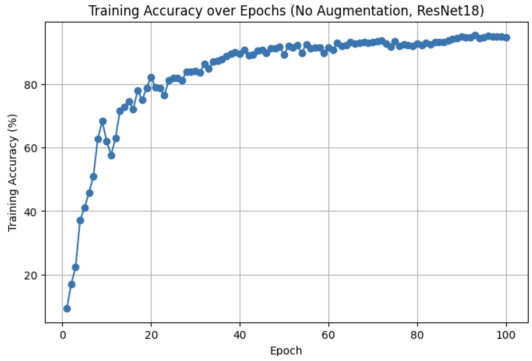

Figure 4. Training set accuracy for 10-image set intentionally overfit to validate setup and set learning rate to 1e-3

context-clashing pairs like "car" and "sofa". A histogram is plotted to show the most popular label pairs. The hypothesis is that only training on the vanilla dataset fails to expose the model to out-of-distribution pairings, which prevents the model from being able to generalize when labelling real world data with unseen pairings. By using this co-occurrence list to filter for rare pairings, we can control how unusual the augmented data is compared to the raw dataset.

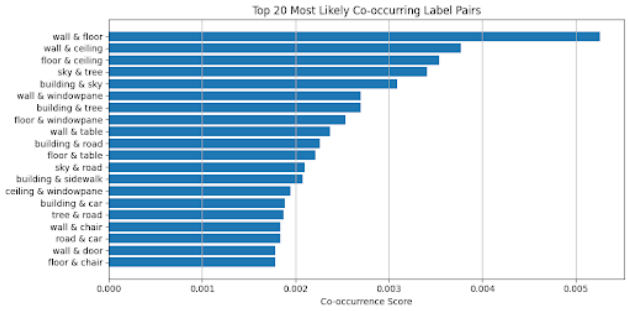

Figure 5. Frequency of label pairs co-occurring in train dataset

Given the size of the dataset, the pairing space is very sparse. This log scale graph shows the co-occurrence likelihood in a histogram, showing that most pairing occur at a near-zero likelihood. Only very select pairs actually repeat across the dataset such as 1 and 4, which are "wall" and "floor".

The co-occurrence list is split into 5 quantiles, with quantile 0 being the rarest pairings, and quantile 4 being the most common pairings. This quantile choice is a hyperparameter when augmenting the data. The second hyperparameter is num_aug, the number of augmented data points. Ablation studies are conducted to find the impact of the augmentation size and rarity of augmentation content on the model learning.

To augment the data, the helper function takes in the original dataset and the specified quantile of the co-occurrence list for that dataset. It generates num_aug new

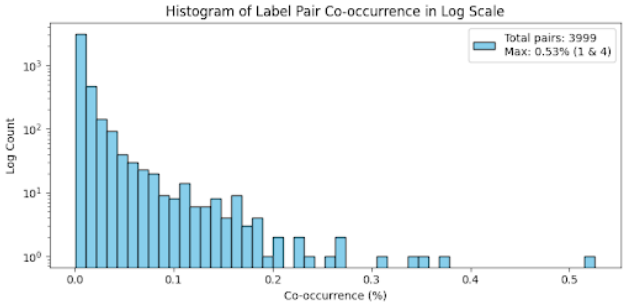

Figure 6. Log scale histogram of frequency of co-occurrence; most pairs are extremely rare

Algorithm 1 Context-Clash Augmentation Procedure

```
Compute co-occurrence matrix of label pairs
Select quantile of rare pairs for augmentation
for i = 1 to num_aug do
    Select label pair (l_a, l_b) from quantile subset
    Find image I_a containing l_a and image I_b containing l_b
    Extract mask M_a of label l_a from I_a
    Copy pixels from I_a masked by M_a onto I_b
    Append augmented image and mask to dataset
end for
```

images by looping over the quantile subset of label pairs, searching for image_a with label_a, and image_b with label_b from the original dataset. Then, it makes a sticker mask from image_a where the pixel label is equivalent to label_a, and copies it onto image_b by replacing the pixels in image_b and its segmentation mask. This new image is appended to the dataset. For computation efficiency the label to image lookup table is created ahead of time.

4.5. Example of Overlaid Synethetic Image

To visualize the synthetic image, a helper function is used to generate a plot showing the side-by-side of the original image and the augmented image as follows. The example image mixes together doctors with a building facade, which adds novel context to the images.

5. Experiments

There were two main experiments to validate the effectiveness of the data augmentation technique exploring how the amount of augmented data and the uniqueness of the pairs used for augmentation affects the validation accuracy.

The training set has 1010 samples. For experiment 1 it was augmented by 10%, 100%, 300%, 500% to 1100, 2020, 4040, and 6060 samples respectively. The PSPNet with ResNet-18 backbone was trained 3 times each with batch size of 8 and epoch of 5. The training loss, optimizer, etc are kept constant as per the Methods section. Overlaying the

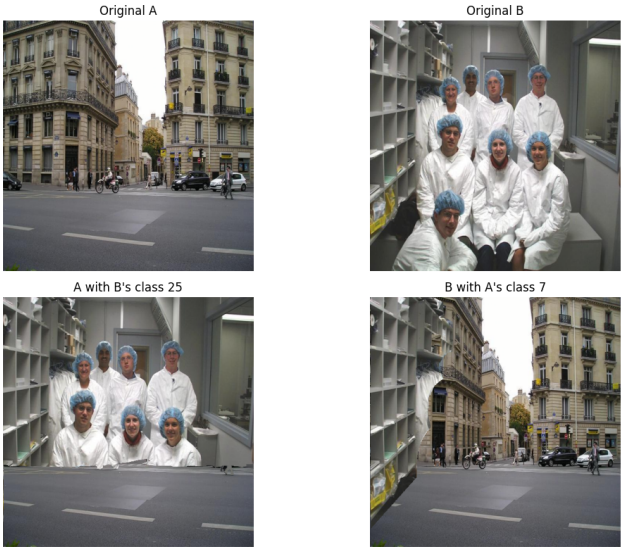

Figure 7. Example of synthetic data with two mixed images

validation accuracy of the models with different augmentation amounts, it can be observed that an augmentation of 300% results in the best performance of the model, with an improvement from 49.88% to 52.65% validation accuracy at epoch 5.

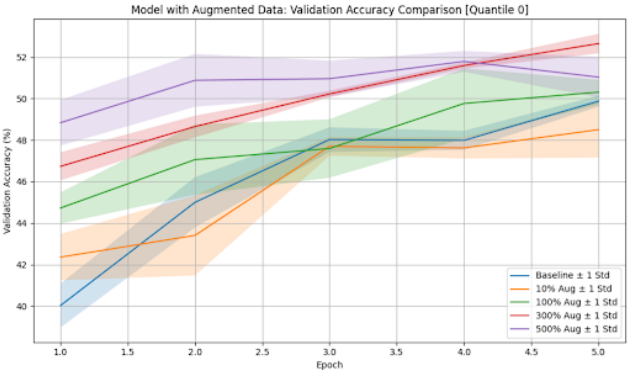

Figure 8. Model validation accuracy varying augmented dataset size

% of Aug Data	Final Validation Accuracy
Baseline	49.88%
10%	48.50%
100%	50.31%
300%	**52.65%**
500%	51.03%

Table 1. Validation accuracy for models trained with different amounts of augmented data. The 300% augmentation yields the best performance.

However, increasing the augmentation to 500% resulted in a slight drop in validation accuracy to 51.03%, despite

a higher training accuracy. This suggests overfitting to the augmented data. This can also be seen in the training versus validation accuracy graphs, where the training accuracy becomes much higher than validation accuracy and validation accuracy begins to stagnate. This is showing how the model is memorizing and overfitting to the training data. Since the synthetic samples were generated by copy-pasting regions within the original training images, many visual features remained redundant. At high augmentation levels, the dataset becomes dominated by these partially repeated patterns, reducing the model's ability to generalize to unseen, real-world scenarios.

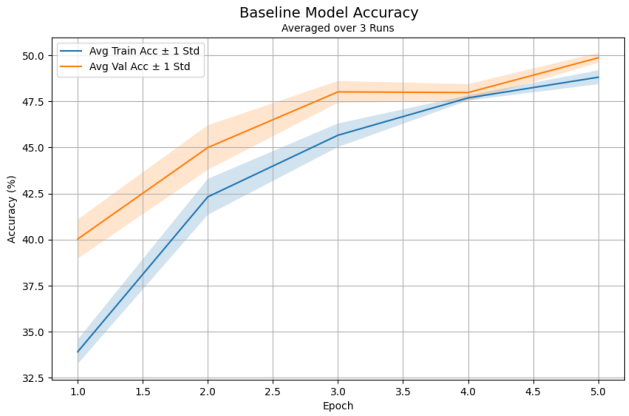

Figure 9. Baseline model training and validation accuracy over 5 epochs

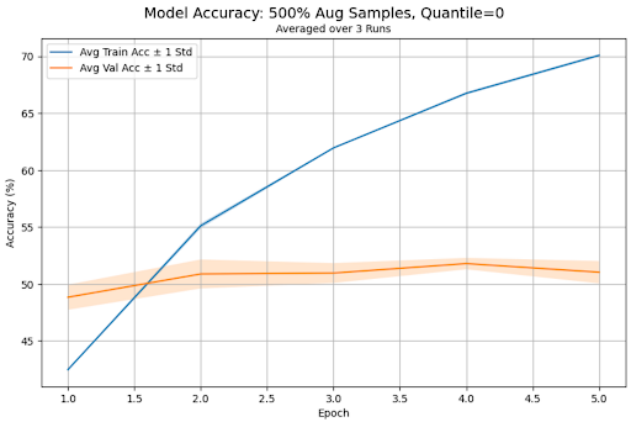

Figure 10. 500% augmented data training and validation accuracy over 5 epochs

This highlights the importance of not just how much but also how data is augmented. Careful tuning of the augmentation scale is essential to avoid overwhelming the model with synthetic noise.

The second experiment fixed the augmentation amount at 100% and instead varied the quantile of the co-occurrence likelihood used to select image pairs for augmentation. These quantiles correspond to how frequently certain image regions are likely to appear together. The aim was to test the hypothesis that combining rare, out-of-distribution image pairs might help the model generalize better.

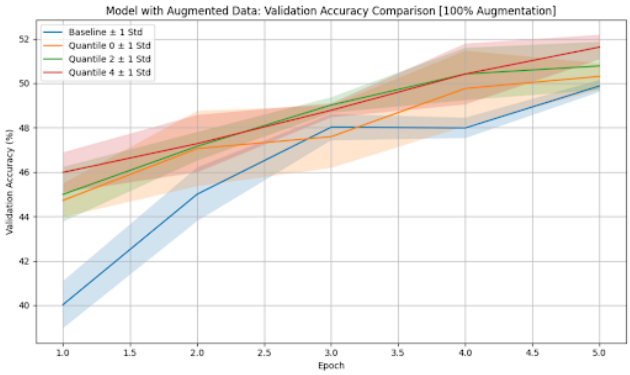

Figure 11. Model validation accuracy varying the quantile of sorted co-occurrence list used to generate the augmented pairs for synthetic dataset

Quantile	Final Validation Accuracy
Baseline	49.88%
0 (Least Likely)	50.31%
2 (Medium Likely)	50.78%
3 (Most Likely)	**51.63%**

Table 2. Validation accuracy for models trained with different co-occurrence quantiles in the augmented data. The highest accuracy is achieved using the most likely co-occurrence quantile.

Contrary to the hypothesis, the best performance was observed when using image pairs from the most likely co-occurrence quantile (Quantile 0). All augmented models outperformed the baseline, but the model trained with the most probable pairings achieved the highest validation accuracy. This result suggests that generating synthetic examples from realistic, high-likelihood combinations helps the model better generalize to the validation set. Since the validation data originates from the same un-augmented source as the training data, it is likely to contain these common patterns. Therefore, targeting augmentation on likely but unseen variations enables the model to learn and generalize the dominant visual patterns more effectively and improve validation accuracy the most.

As discussed in the Alernative Methods section, another hyperparameter originally explored is the sticker size threshold. By limiting the allowable sticker to be at a minimum 20% the size of the original image, we are able to prevent augmentations of only very small sections of the image that could be mistaken as noise. However, compared to the baseline this actually worsened the performance of the model, decreasing the validation accuracy from 49.88% to 48.40%, thus this method was not further explored in fa-

vor of a simplier approach with less hyperparameters. A possibility as to why this method did not improve the model is because only large background objects take up more than 20% of the image, and thus this augmentation technique is not able to produce synthetic data with fine-grained objects and the contextual-novelty it can provide is limited.

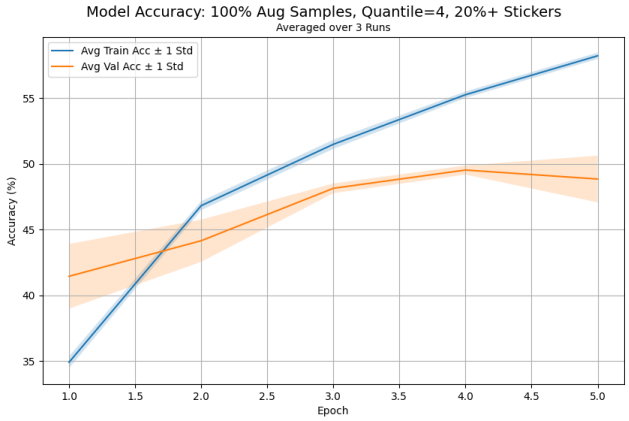

Figure 12. Model training and validation accuracy when imposing a 20% minimum sticker size constraint

6. Conclusion

In conclusion, the experiments demonstrate that carefully controlled data augmentation can improve segmentation-labelling model performance, especially in low-data regimes. Increasing the amount of augmented data up to a point enhances validation accuracy, but excessive augmentation risks overfitting due to redundancy in synthetic samples. Moreover, selectively generating augmented pairs based on common co-occurrence likelihoods proves more effective than focusing on rare or unlikely combinations, as it better aligns with the distribution of the validation data. These findings highlight the importance of balancing augmentation quantity and quality, and of using contextual co-occurrence information to guide augmentation strategies that promote robust generalization.

Some further steps to explore would be to graph a histogram of the correctly labeled pixels and then target the synthetic data creation to the most commonly incorrectly labelled pixels. This add-on to the method should see even greater improvements in the segmentation accuracy by intentionally augmenting the dataset where the model does not perform well to account for a lack of diversity in the dataset in that category. Since this paper has shown potential in using context-aware augmentation to improve model robustness, further modification to isolate and target model classification weakness should improve this technique even more.

7. References

References

[1] G. Ghiasi, Y. Cui, E. D. Cubuk, B. Zoph, T.-Y. Lin, and Q. V. Le. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2928, 2021. 1

[2] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning, 2021. 1